FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Exploring search engine counts in the identification and characterization of search queries

Diogo Magalhães Moura

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Carla Teixeira Lopes

July 26, 2018

Exploring search engine counts in the identification and characterization of search queries

Diogo Magalhães Moura

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: João Moreira
External Examiner: Ricardo Campos
Supervisor: Carla Lopes
July 26, 2018

Abstract

The Web is the greatest source of information nowadays and it's frequently used by most people to seek specific information. People do not search the most efficient way possible. They tend to use a small number of search terms, look into few retrieved documents and rarely use advanced features. This makes it harder of finding the intent and motive of the search and provide the user with relevant and precise documents among the numerous documents available online. This topic is particularly important when the information searched is in the health domain. Health-related searches have gained much popularity. In the U.S. only, 72% of Internet users search in this domain. Being able to provide better results could encourage people to be more participative in managing their health and have a fundamental impact on their quality of life.

The purpose of this investigation is to explore the usage of a semantic similarity measure, called "Normalized Google Distance" (NGD), to enrich and increase the context of a health query. To achieve this goal, we explored, implemented and evaluated several methods for classifying queries into three dimensions: health-related, severity and semantic type. For each of the dimensions, we implemented two types of classifiers: NGD Direct Comparison and NGD with SVM. For the first method, the NGD is determined using a set of terms and compared its value with multiple thresholds. The second method combines Support Vector Machines with the NGD values calculated from a set of terms related with the classification and a set of unrelated terms. It is also crucial to retrieve the search engine counts, used to determine the NGD, in fast and light weighted way. So, we developed several retrieval methods based on API and search engine scrapping from the Bing and Google search engines.

To evaluate our solution, we used Portuguese and English queries, extracted from AOL and Sapo Saúde search engines. Along with this, two datasets for the severity and semantic types classification were built. For evaluating the severity methods in English we constructed a dataset based on a report from the World Health Organization with the most deadly diseases. For the semantic types, we exhaustively extracted all the medical concepts from the AOL health queries by calculating the combinatorics permutations of its terms. Finally, we retrieved the semantic types from the medical concepts.

Regarding the results achieved, The more generic classifications like the health-related method obtained better results than more specific ones, like the severity and semantic types methods. In any case, all the results were considered satisfactory. The SVM with NGD has proven to be the fittest method to classify both health-related and severity dimensions. The semantic types classification both methods obtained similar results. Therefore, we concluded that the NGD is, in fact, a valuable asset in query classification and can be used to improve the context behind a user query.

Resumo

Atualmente, a Web é a maior fonte de informações e é frequentemente usada, pela maioria das pessoas, para procurar informações específicas. No entanto, os utilizadores não pesquisam da forma mais eficiente possível. Têm tendência a usar um pequeno número de termos na pesquisa, analisam poucos documentos recuperados e raramente utilizam métodos de pesquisa avançada. Isto dificulta a descoberta da intenção e do motivo por detrás da pesquisa e consequentemente impede o acesso do utilizador a documentos relevantes e precisos, entre os numerosos documentos disponíveis on-line. Este tópico é particularmente importante quando as informações pesquisadas estão no domínio da saúde. Pesquisas relacionadas com saúde ganharam muita popularidade no últimos anos. Nos EUA, 72% dos utilizadores da Internet pesquisam neste domínio. Ser capaz de proporcionar melhores resultados poderá encorajar as pessoas a serem mais participativas na gestão da sua saúde e ter um impacto fundamental na sua qualidade de vida.

O objetivo desta investigação foi explorar o uso de uma medida de similaridade semântica, denominada "Normalized Google Distance"(NGD), de forma a enriquecer e aumentar o contexto de uma pesquisa de saúde. Para alcançar este objetivo, explorámos, implementámos e avaliámos vários métodos de classificar interrogações em três dimensões: saúde, gravidade e tipo semântico. Para cada uma das dimensões, implementámos dois tipos de classificadores: NGD Direct Comparison e NGD com SVM. Para o primeiro método, o NGD é determinado usando um conjunto de termos e comparado o seu valor com vários limites. O segundo método combina Support Vector Machines com os valores de NGD calculados a partir de um conjunto de termos relacionados com a classificação e um conjunto de termos não relacionados. Foi também crucial recuperar as contagens do motor de busca usadas para determinar o NGD, de forma rápida e leve. Por isso, desenvolvemos vários métodos de recuperação desta contagem baseados em APIs e *scrapping* de motores de busca através do Bing e do Google.

Para avaliar a nossa solução, utilizámos interrogações em português e inglês, extraídas dos motores de busca AOL e Sapo Saúde. Foi necessário contruir dois conjuntos de dados para a classificação de gravidade e tipos semânticos. Para avaliar os métodos de gravidade em inglês, construímos um conjunto de dados baseado em um relatório da Organização Mundial de Saúde com as doenças de maior taxa de mortalidade. Para os tipos semânticos, extraímos exaustivamente todos os conceitos médicos das interrogações de saúde da AOL, calculando as permutações combinatórias dos seus termos. Finalmente, obtemos os tipos semânticos a partir conceitos médicos.

Em relação aos resultados alcançados, as classificações mais genéricas, como o método de classificação em saúde, obtiveram melhores resultados do que as mais específicas, como os métodos de gravidade e tipos semânticos. Em qualquer caso, todos os resultados foram considerados satisfatórios. O SVM combinado com NGD provou ser o método mais adequado para classificar as dimensões relacionadas à saúde e à gravidade. A classificação dos tipos semânticos dos dois métodos obteve resultados semelhantes.

Concluímos que o NGD é uma mais-valia na classificação de interrogações e pode ser usado para melhorar o contexto das pesquisas.

Acknowledgements

First, I would like to I would like to thank my supervisor, Carla Lopes for all the knowledge shared throughout all stages of this thesis. I would also like to thank my dearest friend Inês Carneiro for her patience and support during this 5 long years of academic studies.

Many thanks to everybody that I had the chance to meet during my passage in MIEIC for always being there for me when needed. It was a great experience studying with such a capable and cheerful group of individuals. Thank you for all the great conversations that we shared.

Last, but certainly not least, I must express my very profound gratitude to my family for presenting me with unconditional support and continuous encouragement throughout my academic studies and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you!

Diogo Moura

"Research is what I'm doing
when I don't know what I'm doing"

Wernher von Braun

Contents

List of Figures

LIST OF FIGURES

List of Tables

Abbreviations

API	Application Programming Interface
CHV	Consumer Health Vocabularies
NGD	Normalized Google Distance
QC	Query Classification
SVM	Support Vector Machines
UMLS	Unified Medical Language System
WSD	Word Sense Disambiguation

Chapter 1

Introduction

The development of technologies has brought many benefits to our society. Nowadays more and more people are connected using the Internet. The Web has become the largest source of information and it is frequently used to seek for information. It has allowed users to search almost every topic and became more informed in their topics of interest.

1.1 Context

There is a great deal of health data accessible on the Web and the Internet has become a major source of health seekers.

A survey conducted in the U.S in 2013 shows that from the 85% of U.S. adults that use the Internet, 72% have looked for health-related information within the past year [Fox11]. A similar survey conducted in the E.U., concludes that six out of ten Europeans search for this kind of information and 75% of them agree that the Internet is a good source for this kind of information [HSB].

In the U.S., 77% of health consumers start the search in a standard search engine like Google or Bing, only 13% begin their search in a specialized search engine for health subjects, 2% in general information websites, like Wikipedia, and 1% in social networks. The other 7% of the respondents answered "Other", "Don't Know" or refused to answer [Fox11]. In the case of Europe, 82% to 87% of health consumers start their search in search engines [HSB].

Despite the fact that nearly 90% of people who looked for health information online considered their search results as successful [HSB], studies concluded there are still a few problems regarding this type of searches. A clear example of one of these problems is the self-diagnosis process performed through the search of symptoms in the search engines. The work done by Ryen White and Eric Horvitz [WH09] show that Web searches have potential to increase anxieties in people that have little or no medical training. The other 10% of health seekers that are not satisfied with the health-related information found on the Internet noted the following problems: unreliable information, excessive commercial oriented data, lack of details and not tailored to their specific needs.

1.2 Motivation and Objectives

People are spending more and more time seeking and retrieving information. Their interactions with search engines are short and limited [SWJS]. They tend to use few search terms, few modified queries, look at few retrieved web pages and rarely use advanced search features. A small number of terms are used with high frequency when there are a lot of better unique terms. Nevertheless, search engines still present reasonably good results for what the user is seeking for. The problem is going beyond the explicit terms entered by the user. Is finding the intent and motive behind that search, refine the results presented to the user, order the relevant documents according to the search intentions and present more appropriate results.

In the health domain, there are some significant problems in query formulation [TL07]. This mainly occurs because of the differences in knowledge and comprehension in the medical domain. For example, a person with less knowledge of medical vocabulary would search for 'heart attack', while the technical term is 'myocardial infarction'. Improvements in information retrieval in the health domain could encourage people to be more participative and informed in this field which could lead to an overall improvement of people's health.

Therefore, the goal of this dissertation is to evaluate if the Normalized Google Distance can be used to enrich and increase the context of a health query. So, we will explore, implement and evaluate several methods for classifying queries into three dimensions. The first method classifies a generic query a being or not a query seeking for health information. The following dimensions can only be applied to queries already established as health queries. The second classification concerns the severity associated with the query. The third method classifies a query as their semantic type. We will classify the query into one or more of the following types: Finding, Disease or Syndrome, and Therapeutic or Preventive Procedure.

1.3 Document Structure

Besides the introduction, this document contains six more chapters. In Chapter 2, it is described the background information and state of the art on Semantic Search, Query Classification and available resources. Chapter 3 presents the problem, solution, and the methodology. The following three Chapters, 4, 5, 6 describe the three types of classifiers and the results obtained. At last, Chapter 7 refers the conclusions and the future work.

Chapter 2

Background and State of the Art

This chapter will describes the current state of Semantic Search and Query Classification. Section 2.1 defines Semantic Search, types and techniques that it uses. Next Section 2.2 will explain Query Classification, most common techniques and work done in this field. Lastly, Section 2.3 describes the existing resources at our disposal.

2.1 Semantic Search

Semantics is the study of the meaning and logic of words. When applied to search, it attempts to find the intent of the searchers and the contextual meaning of the search terms. In search engines, semantics search aims to improve the search accuracy and provide the user with results more adjusted to their intent.

Guna et al. [GMM03] defined two different types of searches: Navigational Searches and Research Searches. In Navigational Searches the user intends to find the query terms in the documents. On the other hand, on Research Searches, intends to find the object or entity associated with the searched terms.

Major search engines have included the semantics in their search to provide the user with better results. The following technics have been used to insert semantic into search:

- Word Sense Disambiguation -– Word Sense Disambiguation (WSD) is finding which meaning of the word is used in the phrase when that word is ambiguous. WSD is a difficult problem to solve since most of the user queries are short and contains multiple generic terms [SWJS]. Several approaches have been explored like dictionary methods, unsupervised learning but the most successful are supervised learning methods.

- Location as Context -– The location where the user is making the search can be used by the search engine to obtain better results. For example, if the query is “weather” then the search engine should be able to provide the user with the weather forecast based on the user location

- Current trend -– The search engine should be able to refine the results based on the current trends and news

To increase the semantic value and enhance its search engine results, Google developed a knowledge-based resource called the Knowledge Graph [GKG]. This huge graph contains millions of entities and relations between them and over 70 billion of facts. When searching in Google for the term "michael phelps", the engine can successfully associate the search with the entity and present to the user that Michael Phelps is a swimmer, his biography, photos and related people apart from documents containing the search terms. This is possible due to the Knowledge Graph.

Grimes [Gri] drafted a list of 11 approaches that join semantics with search. The "Two + Nine Views of Semantic Search" list contains functionalities that can be associated with Semantic Search:

- Related searches/queries – The engine proposes searches that are in some way similar to the entered search. The "Did you mean feature", found in Google, falls into this type.
- Reference results – Search engine returns materials, like maps, images or videos that define the searched terms.
- Semantically annotated results – Highlighting text features like entities that are semantically close to the search terms.
- Full-text similarity search – Search engine examines all the words in every stored document as it tries to match search criteria.
- Search on semantic/syntactic annotations – The user classifies the keywords according to a synthetic role or meaning and the search engines tries to match documents according to those rules.
- Concept search – Search based on relations between concepts. This relation can be obtained using taxonomy or statistical co-occurrence techniques.
- Ontology-based search – The engine extracts semantic meaning from the user's query by means of ontology concepts. The relations between concepts are defined by ontologies, which are more complicated than those by concept search.
- Semantic Web search – Similar to Ontology-based search but the returned data is structured and queryable.
- Faceted search – Searches using facets provide the user to refine the query into predetermined categories.
- Clustered search – Similar to faceted search but without predetermined categories. This technique extracts the meaning from the retrieved documents of the search.

- Natural language search – Greatly used in voice search, the user formulates the query using everyday language and the engine creates a semantic representation of the query.

As stated on Concept Search, co-occurrences can be used as a technique to extract semantic meaning from queries. Various researches were made in this field using co-occurrences as a semantic measure. For example, Nunes et al. [PDC+] developed a solution for discovering entity relations using entity co-occurrence on the Web with the help of search engines. And also, Bullinaria et al.[BL07] concluded that a very simple method of co-occurrence statistics can be used to extract word meanings.

2.2 Query Classification

Information retrieval (IR) "is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)" [MRS09]. Query Classification (QC) is a category of IR that is focused on classifying web search queries. Predicting the category of a query is useful to provide more accurate results and show related documents according to the query category. Besides improving the documents to be presented to the user, Query Classification can also be used to redirect the user to a more focused search engine based on the topic predicted. It is the case of Metasearch engines. This type of search engines does not have its internal information on web pages. Instead, they really on third-party engines to provide them with results. By incorporating query classification mechanisms, they can query specific search engines based on the category of the search and put more emphasis on these results. For example, if a user searched for "Honda", Metasearch engines can show first the results retrieved from a search engine specialized in automobiles and only then the other results. One more applications of Query Classification is in advertising. By classifying what the user is looking to, the advertisement presented to the user can be personalized according to the context of the search.

2.2.1 Techniques

Multiple strategies can be found in the literature related to Query Classification. This process usually involves two phases to successfully associate a category with a query, as shown in Figure 2.1.

As users tend to use few terms and ambiguous terms in their searches [SWJS], the process of classifying is hard. Therefore the first phase in QC consists of enriching the terms with more information related to them. Some authors use information from retrieved documents. For example, Shen et al. [SPS+05] used the category of the top ranked documents returned by the search engine. Search logs can also be used in this task. Cao et al. [CHS+] took advantage of refinements of the search queries that users made during their search session and D. Le and R. Bernardi [LB12] of these links the searchers clicked.

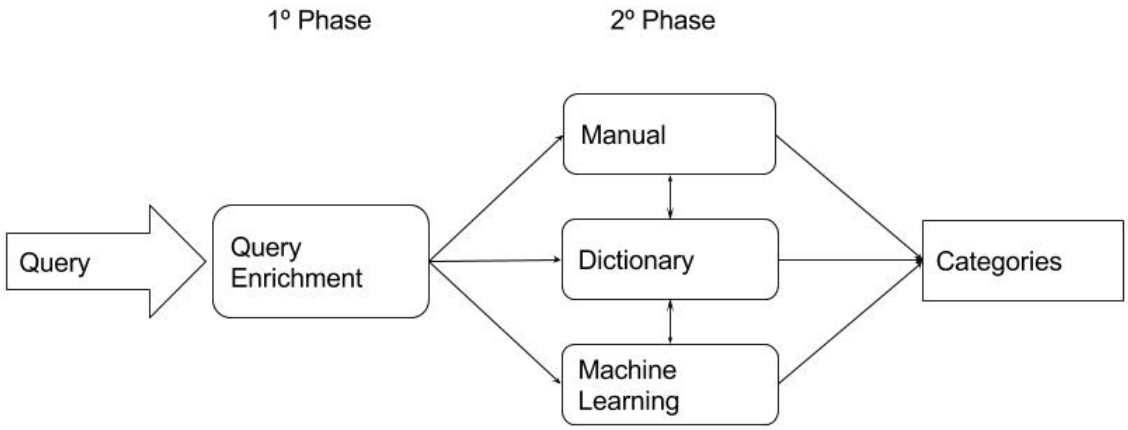

Figure 2.1: Process of Query Classification

The second phase is the classification itself. The first method found in the literature is the manual classification. It consists of human experts in the field manually labeling all the datasets. It is usually done by more than one expert to provide consistency and validation of the given labels.

A dictionary is a collection of words regarding a specific topic. To recognize the topics in the query, a matching against the dictionary terms is performed. There are multiple ways of string matching against a dictionary like exact matching, approximate matching, and phonetic matching. On exact matching, the words need to be exactly as specified in the dictionary for a positive match. However, some flexibility may be required. For example, stemming or replacing special characters can be applied to the words before matching. Approximate matching, also known as fuzzy matching, calculates edit-distance or Jaro-Winkler's metric between two words, comparing the differences between them. The matching is successful below a given threshold value. Lastly, phonetic matching compares words based on what they sound. One of the most used algorithms is the Soundex. In a simplified way, it encodes to the same internal representation so that homophone word have the same representation and can be easily compared.

The base idea of machine-learning methods is to train computational models without explicit being programmed so that they can make predictions on the given data and learn from the mistakes made along the way. On the field of Query Classification, machine learning is applied to make predictions on the characteristics of a given query. Machine learning models can be classified in three categories, depending on the used data: supervised learning uses labeled data on the training phase to generate a function that maps inputs to desired outputs; unsupervised learning does not require annotated data and applies appropriate functions to infer data; semi-supervised learning mixes the two previous approaches. Several studies tried different techniques and algorithms in this field, but Support Vector Machines seemed to be the most widely used.

It is also possible to combine these techniques, mainly dictionary and machine learning, into a single classifier. Beitzel et al. [BJF+] and Shen et al. [SPS+05] used this approach in their work and proved to increase performance.

2.2.2 Evaluation

Regarding evaluation of the produced systems, metrics were defined and standardised. Although there are some slight differences in the evaluation across different researchers, they are all based on the same base concepts: precision, recall and F-measure. These metrics are obtained through the comparison of a manually classified query against an automatic classified one. First, it is necessary to understand the following concepts:

- True Positive (TP) — the topic is present in both queries.
- True Negative (TN) — the topic is not present in none of the queries.
- False Positive (FP) — the topic is present in the automatic classified query,but not in the manual.
- False Negative (FN) — the topic is not present in the automatic classified query, but it is the manual.

Precision is the ability of a system to present only relevant items, and is formulated as:

$$Precision = \frac{TP}{TP+FP} \tag{2.1}$$

Recall, often called sensitivity, measures the ability of a system to present all the relevant items, and is formulated as:

$$Recall = \frac{TP}{TP+FN} \tag{2.2}$$

F-measure, also called F1 score, is the harmonic mean of precision and recall and is formulated as:

$$F-measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.3}$$

Another relevant measure when doing binary classification is the Cohen's kappa coefficient. It measures the inter-rater agreement for categorized items and takes into account not only the percentage of agreement calculation, like Precision but also the possibility of random agreement. For binary classification, it is calculated using Equation 2.4, where po is the observed agreement and pe is the agreement by chance. This measure can be hard to interpret and there isn't a consensus on it's meaning. Some authors [LK77] characterized values below 0 as bad, 0-0.20 slight, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial and bigger than 0.81 as almost perfect. Others [FLP03] classified kappa values below 0.40 as poor, 0.40-0.75 as good and 0.75 as excellent.

$$Kappa = \frac{po - pe}{1 - pe} \tag{2.4}$$

2.2.3 Related Work

In the topic of query classification, several studies were conducted over the years. A study focused on analysing the web queries made by the public to the Excite search engine conducted a manual classification on a random sample of 2,414 queries [SWJS]. The classification had 11 categories and the category that came on top was "Entertainment, recreation" (16.9%) followed by "Sex, pornography, preferences" (16.8%). Only 6.8% of the queries belonged to the health field. A similar study focusing on health queries also took a manual approach classifying queries from "Excite", "AlltheWeb.com" and "Ask Jeeves" search engines [SYJ+04]. This kind of classification is very expensive, can take a long time to be done and represents a tedious work for the human classifiers. That is why several automatic methods have been proposed.

An example of these automatic methods can be found in a study done by Beitzel et al. [BJF+, BJL+07]. The researchers developed three different classifiers, Exact Matching, Supervised Machine Learning with Perceptron using an Margins algorithm, and Selectional Preferences, to build a high-recall classification system for general web queries. The exact matching classifier looks for the query in a database of manually classified queries. Based on the labeled queries from the previous approach, they created a linear binary classifier for each category using all queries in a given category as positive examples and all queries not in that category as negative examples. Selectional Preferences is based on "the tendency for words to prefer that their syntactic arguments belong to particular semantic classes". They concluded that a combined approach of this techniques achieves approximately 20% higher recall than any single approach.

In 2005, the Data Mining and Knowledge Discovery competition (KDD Cup) focused on this topic. The challenge was about automatically classifying 800000 internet user search queries into 67 predefined categories. The winning solution [SPS+05] achieved an average of 94,4% in the F1 value. In their approach, they developed synonym-based classifiers, a statistical classifier in which Support Vector Machine (SVM) is employed and two ensemble classifiers which improve the classification performance significantly more than previous classifiers. Before classifying the queries they first performed a method for query enrichment [SPS+06]. To enrich the queries, they search for their terms in three search engines and classified them according to labels of returned documents. Since the labels from the returned documents are different from the KDD categories, they developed a synonym-based classifier for each search engine to map the categories. This classifier had a low-recall so they developed a classifier based on SVM. To represent a query they capture information, like titles, snippets, URLs terms, from the top N results when searching the query on a search engine. This solution was later improved by using a taxonomy-bridging classifier [SSYC06]. This classifier is used to map user queries to the target categories via the above intermediate taxonomy. By using this technique combined with the winning solution they were able to improve the precision by 9,7% and the F1 value by 3,8% .

D. Le and R. Bernardi developed a query classification system in the art, cultural and history domain [LB12]. They showed how click-through links, the links that a user clicks after submitting a query, can be useful to enrich the query. They enrich the queries by extracting click-through

image's title and keywords. To build the dataset they mapped each query to its click-through information to extract the category associated to the corresponding image. They build an SVM model using this dataset. The result from this study has shown an increase in the performance of QC when enriching the queries with click-through information.

All the above approaches use individually queries to try to understand the user intent by that search. It is also possible to use multiple queries as shown by Cao et al. [CHS+]. In their approach they used neighboring queries, queries that result from the refinement of the search by the user and their corresponding clicked results in search sessions as the context information. Then created a classifier by using conditional random field models. They used Excite search engine to extracted the user queries during their search session and manually classified each session into the taxonomy of the KDD Cup'05. They concluded that this approach outperforms other that do not leverage from search context. However, it cannot solve individual queries as well as others.

Agrawal et al. [AYKZ11] developed two approaches simpler than methods using Machine Learning methods. The first approach, category enrichment method, extracts the metadata from existing documents in the Web when searching for the category in a commercial search engine. The extracted metadata such as document title, keywords, and page content is used to build a search index in a commercial search engine called Sphinx. Then the query is classified against this index. The second approach can be seen as the opposite of enrichment. The reductionist approach reduces the query to few central terms. The query is label based on this central terms instead of using the whole query. The first method has a high recall but low precision, while the second has high precision but low recall.

E. Diemert and G. Vandelle [DV09] developed an automatic categorization system that doesn't require previously label queries to train the models. Instead, they followed an unsupervised learning scheme. The model is stored in the form of a concept graph. The nodes correspond to concepts and the edges cross-reference between concepts. This graph is being dynamically built from Yahoo! search query logs through a mining process. The system was tested using the KDD Cup 2005 dataset where it achieved a lower precision than the winning solution but a better F1 score. It is also noted that this system has been successfully deployed in production on Yahoo! Search UK. In fact, the results of the classifier influence the engine query execution plan to display different layouts and rich content to the user.

Most recent works focused on using word embedding in QC. The researchers developed a system based on the Word2Vec algorithm named TOWE [YHH15]. This system achieved the 95.73% in terms of Precision, 97.79% in terms of F1.

There are also papers focusing only on Query Enrichment like "Wikipedia-based semantic query enrichment" [ABC13]. They took advantage of Wikipedia to develop a semantic enrichment method. The method enriches the query by adding the title of related Wikipedia articles. For example, the query "Last Supper" was enriched by three articles with the title: "Jesus", "Crucification" and "Twelve Apostles".

In the health domain, there are not many targeted approaches besides the work of Eysenbach and Kohler [EK03] and Carla Lopes and Cristina Ribeiro [LR14]. Eysenbach and Kohler [EK03]

proposed a method to automatically classify search strings as health-related based on the proportion of pages on the Web containing the search string plus the word "health" and the number of pages containing only the search string. Carla Lopes and Cristina Ribeiro [LR14] developed three different methods to identify and classify health queries. The first method is based on the work described previously, but instead of only using Google for the queries counts, they combined the counts of Yahoo and Google search engines. The following methods take advantage of the Consumer Health Vocabulary (CHV) to classify the queries. One of these methods uses an exact matching between the terms in the query and the CHV. If the query has at least one term that is also in the health vocabulary then the query falls in the health category. The last method differs from the previous one, since it produces a continuous output.

Table 2.1: Summary of work in Query Classification

Authors	**Data Sources**	**Techniques**	**Algorithms**
Spink et al.	Excite query logs	Manual	-
Spink et al.	Excite, AlltheWeb and AskJeeves search engines	Manual	-
Beitzel et al.	AOL query logs	Dictionary Machine Learning	Exact Matching Perceptron Selectional Preferences
Shen et al.	KDD CUP '05 Dataset	Dictionary Machine Learning	Synomy Matching SVM
Le et al.	Bridge Man Art query logs	Machine Learning	SVM
Cao et al.	Excite query logs	Machine Learning	Conditional Random Field
Cao et al.	Baidu and Sogou query logs	Machine Learning	Word2Vec
Agrawal et al.	Logs of two non-specified search engines	Dictionary	Approximate Matching Synonym Matching
Diemert et al.	Yahoo! Search query logs KDD CUP '05 Dataset	Machine Learning	Knowledge Based Search using a concept graph
Eysenbach et al.	Excite and AlltheWeb query logs	Dictionary	Co-occurences
Lopes et al.	AOL and Sapo Saude query logs UMLS CHV	Dictionary	Co-occurences Approximate Matching

To sum up, Table 2.1 shows there is a great variety of techniques and algorithms are used to classify queries. Query enrichment proved to be beneficial and improve the performance of the classifiers. However, finding label queries is a significant obstacle. Popular search engines restrict access to their query logs, and most of the work here presented have to use the same data sources.

2.3 Available Resources

2.3.1 Unified Medical Language System

Having knowledge bases, like databases and dictionaries make the process of QC much simpler eliminating the need to use sophisticated methods based on Machine Learning. Unified Medical Language System (UMLS) [uml09] is an enormous knowledge base that aggregates multiple

dictionaries and vocabularies in the medical domain. It has been referred as being "probably the most comprehensive ontology in healthcare" [NM04]. It is composed of three main knowledge sources, which ease the developments of applications in the biomedicine and health information field, which are:

- Metathesaurus - a large, multi-purpose, and multi-lingual thesaurus that contains millions of biomedical and health related concepts, their synonymous names, and their relationships. It holds over 1 million biomedical concepts, 5 million concept names and nearly 200 controlled vocabularies and classification systems.
- Semantic Network — semantic types and semantic relations that provide a consistent categorization of all concepts in Metathesaurus and relationships between them. It reduces the complexity of Metathesaurus and makes navigation more accessible since the concepts are grouped into small and coarser-grained semantic types.
- SPECIALIST Lexicon — provides the lexical information needed for the SPECIALIST Natural Language Processing Tools. It also includes commonly occurring English words and biomedical vocabulary.

Due to its size and complexity, the UMLS contains some errors and inconsistencies [GMX+09]. These errors include ambiguity and redundancy once the same concepts can be defined in multiple vocabularies, child and parent relationship is not consistent with the concepts and lack of ancestors.

The UMLS provides four ways of accessing its content. The Metathesaurus browser is a web interface that allows users to query the different vocabularies present in its database and search specific concept unique identifiers. It also provides a way of navigation, in a tree-like style, through all the sources and retrieve the correspondent concepts, semantic types and relations between other concepts. It is also possible to download the entire UMLS and generate scripts to import the data to different relational datasets. The last two are HTTP APIs, one using REST and other using SOAP. They are both relatively simple to use, with few endpoints and restrictions. On the other hand, the authentication is more complex and out of the ordinary. It requires an API key obtained in the UMLS platform, following a Ticket-Granting Ticket which is only valid for 8 hours and finally a one-time use service ticket. By obtaining this last ticket, it is possible to access UMLS information.

2.3.2 Consumer Health Vocabularies

Health consumers have difficulties finding and understanding health information. The gaps in knowledge are the main contributor to this factor. In fact, studies have found that using medical terms provide better results than its related consumer terminology [PMS+01]. Thus, there was a need to map those two terminologies in a vocabulary named Consumer Health Vocabulary (CHV). The building of the CHV was a complex task that required steps like obtaining health consumer terms, developing automatic methods and manual reviews by medical personnel.

Multiple approaches were developed like analyzing queries from health information sites and identifying consumer health displays [ZTC+05] and mining terms from community-generated text [VMHZ14].

An open access collaborative consumer health vocabulary was created in 2006 [ZT06]. This CHV maps consumer health terms with Unified Medical Language System. This system contains a set of health and biomedical vocabularies and medical standards. Currently, this CHV is a part of the UMLS.

A study conducted in 2007 found that some terms in the open access collaborative CHV do not map to UMLS concept which affects the process of consumer health vocabulary building [KSD+08].

In what concerns the Portuguese language there aren't any CHV available. There is a project proposal to develop a CHV for the Brazilian Portuguese [KSD+08] language that should have started in the year of 2016. The first phase of the project pretends to extract terms and connect them using automatic techniques, then validate the concepts using human reviews and lastly store the data in a standard model to facilitate data exchange.

2.3.3 Datasets

There are three datasets that can be used either to evaluate the results from the methods or to use as training data.

The AOL Searches dataset was obtained by using Automatic Query Classification via Semi-supervised Learning in the users queries to the AOL web search service [BJF+]. This set contains 1,197 health queries from a total of 23,780 queries in English (EN) classified in one of the following categories: autos, business, computing, ent, games, health, holidays, home, misspell, news, org, other, pf, places, porn, research, shopping, sports, travel, url.

The Sapo Saude data set is a collection of 1,522 queries manually classified by medical students in Portuguese (PT). This data set contains queries submitted a health search engine called Sapo Saúde. The dataset contains the following fields:

- query - terms searched by a user in Sapo Saude search engine
- lang - Language of the query. If the value is PT then the query is in Portuguese. If it is 0 then the query is in another language.
- sum_health - The number of labelers who consider the query to be of in the health domain.
- sum(terms) - The number of labelers who consider the query contained a medico-scientific term.
- sum(severity) - The number of labelers who consider the query to be about a severe medical condition.

The KDD Cup 2005 dataset is compilation of 800 manually classified queries by 3 different human labellers. Each queries has multiple topics associated being the most relevant "Living\Health

& Fitness". From the 800 queries, 69 are considered health queries. In Table 2.2 are a sample of the queries present in the three datasets.

Table 2.2: Example of queries present in the KDD, AOL and Sapo Saude datasets

KDD	**AOL**	**Sapo Saúde**
permenant abdominal gas remedies	suicide and sleeping pills	doenca de alzeimer
cheap international airfare	kitchen designs	neuropsicologia
divinity candy	positions	divorcio
new zealand clothes	mortgage underwriter	dicionario portugues
uk telephone directory	sunquest cruisez	receitas vegetarianas

Chapter 3

Problem and proposed solution

This chapter presents a more detailed description of the problem, proposed approaches, implementation details that need to be taken into consideration and some evaluation metrics for measuring the performance of the methods.

3.1 Problem

As seen in Section 1.2, user's queries are short and ambiguous. These characteristics make the task of finding the user's search intents hard. Query Classification has been widely studied for this purpose. Successfully mapping general user queries to predefined categories can bring improvements in the efficiency and effectiveness of general web search. Most of the work done in this field has been towards classifying general queries but would it be possible to classify queries in a target domain like health?

This dissertation aims to explore the use of a semantic similarity measure called Normalized Google Distance (NGD) to classify user queries. This classification will focus on the health domain in both Portuguese and English language.

3.2 Solution

To analyze the feasibility of using the NGD in classifying health queries, we implemented several methods to classify queries into the following dimensions:

- Health-related - The query belongs to the health domain, or it does not.
- Severity - The query is considered to be severe regarding its underlying medical context.
- UMLS Semantic Types - Classification according to some of the types defined in the UMLS.

3.2.1 Normalized Google Distance(NGD)

The Normalized Google Distance (NGD) [CV07b, CV07a], proposed by Rudi Cilibrasi and Paul Vitanyi, is a semantic similarity measure based on the number of results returned by a search engine, in this case Google, for a given number of terms. The Normalized Google Distance is derived from the earlier Normalized Compression Distance. This latter method is feature-free. This means it doesn't analyze the object looking for particular features, instead it analyzes all features simultaneously and determines the similarity between them according to the most dominant feature. It only uses the name of an object and obtains knowledge about similar ones by using the information generated by the millions of web users.

3.2.1.1 Definition

The Normalized Google Distance between two search terms x and y is

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}} \tag{3.1}$$

where:

- The function f(x) is the number of hits returned by the search engine when searching for the term x
- The function f(y) is the number of hits returned by the search engine when searching for the term x
- The function f(x,y) is the number of hits returned by the search engine when searching for the combination of both x and y using an AND operator.
- M is the total number of pages indexed by that search engine

3.2.1.2 Properties

The main properties of NGD are:

1. The NGD values are between 0 and ∞. The NGD can be ∞ when two terms never occur together but do occur separatly. In this case we can consider the two terms dissimilar. In practice, the NGD range is between 0 and 1.
2. The NGD is always nonnegative and the NGD of two similar terms is 0.
3. The NGD is *scale-invariant*. This means if the number of indexed pages increase, so does the number of pages containing the two terms.

3.2.1.3 Work using NGD

Cilibrasi et al. [CV07b] developed three different methods, in order to demonstrate the applicability of the measure. The first method they used unsupervised learning in the form of hierarchical clustering to classify colours, numbers and Dutch Painters, then supervised learning using Support Vector Machines and lastly matching using correlation.

For the investigation at hands, the most exciting method was the second one. The authors conducted three different experiments using SVM Learning to classify terms as emergencies, prime numbers and WordNet categories. To represent the terms into training vectors, they defined several terms, called *anchor words*. Half of these anchors are directly related to the classification under consideration and the other half isn't. Them they calculated the NGD value for each pair of training term/anchor. The result is the training vector used to train the SVM model. The training data was particularly small in all the them, but the result was positive. The NGD had a success rate of 87.25% with a variance of ≈ 0.01 and a standard deviation of ≈ 0.11. It was rare to find agreement lower than 75%.

Evangelista et al. [EKH09] studied the value of NGD between random words selected from news articles. They concluded that the expected value was 0.7 and and proposed a new equation. This equation is a recalibration of Equation 3.1 divided it by 0.7.

The NGD has also been used, with success, to address the problem of discovering mappings between concept hierarchies [GtKAvH07].

Other authors [PZ06] also used NGD in two clustering methods, Spectral Clustering and Semidefinite programming, to decomposed a list into semantically related groups.

3.2.2 Support Vector Machines

Support Vector Machines (SVM) is a machine learning algorithm mostly used in binary classification. The idea behind is to plot each data item as a point in an n-dimensional space. The classification is executed by finding the hyper-plane that splits the two classes. Most of the times the hyper-plane is chosen by the largest separation between points from the two classes. When the data cannot be clearly separated by a linear classification line, and there is some overlap in the data plot, SVM provides three tuning parameters. The kernel is the first parameter. It provides support for nonlinear classification. The Regularization parameter or C parameter is the cost of misclassifying a training point. Large values of C tend to provide a better classification but take more time to compute the hyper-plane. Alternatively, low C values will take less time but misclassify more points. The last parameter, Gamma, represents the range of influence from the training points in the separation line. For example, on high Gamma values, only nearby points are considered. On low Gamma values, far away points are also considered. The best combination of C and Gamma is can be picked using a grid search through all the possible values. Cross-validation is usually the selected method to picked the optimal values for C and Gamma. SVM can also be applied in multiclass classification. This is usually done by splitting a multiclass classification into several binary classifications by either using a *one-versus-all* or *one-versus-one* strategy. The first strategy

involves training a single classifier per class, considering elements from that class as positive and the rest as negative. This approach needs to output a confidence level besides the class label. The classifier with the highest output value assigns the class. The second strategy consists of creating a binary classifier for each pair of classes and creating a voting system. In the end, the class with more votes is considered the predicted class.

3.3 Metodology

3.3.1 Estimating the number of webpages containing a set of terms

Being able to retrieve the number of results quickly is a crucial point of this investigation. Calculating the NGD for a single query can take up to four calls to the search engine. Two for the individual term, one for the combined term and one for the M value. Three different retrieval methods were developed. One method based on REST API calls and two on web scrapping search engines. The methods take advantage of the two most popular search engines, Bing Search and Google Search [sep]. Node.js was chosen as the primary framework used for this retrieval method due to its great asynchronous capabilities and ecosystem. All the retrieval methods presented are implemented in separate scripts according to the method used, the search engine and type of classification.

3.3.1.1 API

The most trivial way to query a search engine is through their REST API, but the search engines impose some hard limits in their quota. For example, Google Custom Search API allows up to 100 API calls per day for free, 100 queries per 100 seconds and has a total maximum of 10,000 queries. On the other hand, Bing Search API also offers a free trial with 30,000 API calls per month, up to 3 per second. Yahoo has also considered a possible candidate, but their API was discontinued at the end of March 2016.

The developed script reads Comma-separated values or Semicomma-separated values file. This file usually contains the queries, a field for the search engine counts of the correspondent query, usually named count-simple and a field for the search engine counts with the query combined with other words, for example, count-health. Next, the file is validated and divided into a chunk with queries to be made. Then, the API is called with this chunk of queries with a configurable number of concurrent requests. Finally, this chunk is inserted back to the whole group of queries and written in the correspondent file. Diagram 3.1 represents the process here described.

3.3.1.2 Web Scrapping

Since the API method is a bit limited in terms of available quota, we implemented two new methods to mitigate this problem using search engine scrapping. However, search engine scrapping introduces its own problems, like blocking and limiting the number of requests that come from the same IP or with the same cookie, forcing captcha resolution before providing the webpage

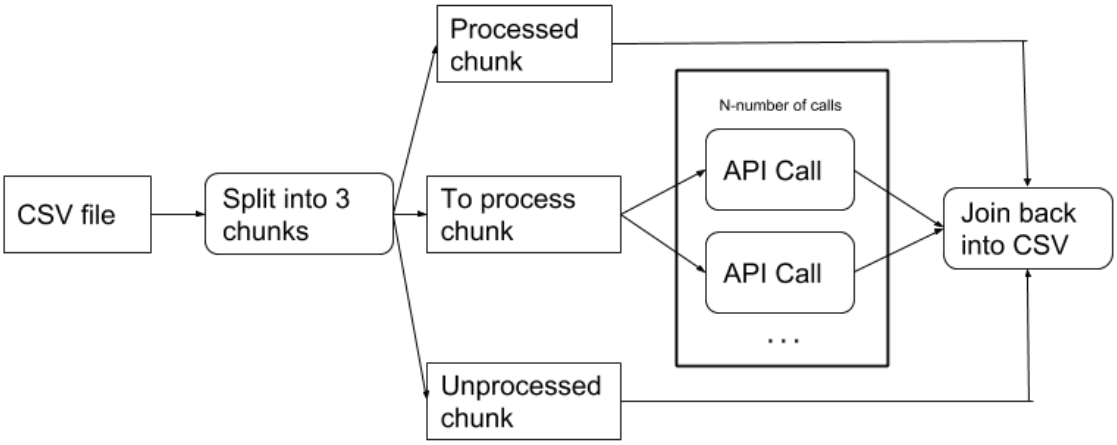

Figure 3.1: Diagram of the script for retrieving search engine counts through the API

content and introducing same small changes in HTML. The following problems were found when scrapping Google, Bing, and Yahoo! Search Engines:

- Google - Captcha when requesting over 800 queries
- Bing - Provided an incomplete page and immediately redirected to the correct one
- Yahoo! - Blocked when making multiple requests

The first method created replicates the GET requests made by the browser when users search for specific terms. Using this method in Google and Yahoo! Search Engines is impossible due to the problems previously mentioned. However, in Bing, this worked somewhat fine apart from the random incomplete pages and the language definition. Bing may returns different results according to the language defined by the users, for example, search *liiinois driver book* with the engine language set to English and region United States of America it returns 90,500 results but when it is set to Portugal and Portuguese language, it only returns 68 results. Working around this problems would be a hard task using this method. It would involve analyzing the cookies set by Bing and tune its parameters. For these reasons, a second method was developed.

The second method takes advantage of a tool called Puppeteer to emulate the behavior of a first time user in Bing. Puppeteer is a library that provides control over a Chromium instance running in the background. The following steps are always performed for each query:

1. Generated new Chromium instance
2. Create a new page
3. Navigate to *bing.com* address
4. Change the language and region to English if needed. Default is Portuguese
5. Simulate user typing the terms in the query

6. Simulate user pressing the ENTER key
7. Wait for the results page to be completely generated
8. Retrieve the number of results by query the HTML element

This method easily overcomes the problems of the previous method and has a lower error rate and is the default method when queries needed to be made to Bing.

3.3.2 Classification

In each of the three dimension classifications, two different methods were defined, implemented and tested.

The first method, called NGD Direct Comparison, is based on the premise that NGD values close to 0 mean that the terms are similar at the semantic level and values greater than 1 are quite dissimilar. It makes a direct comparison between a set of predefined thresholds. If the calculated NGD value is lower than that threshold, then the query is positively classified.

The second method combines the Normalized Google Distance with Support Vector Machines as previously proposed by Rudi Cilibrasi and Paul Vitanyi [CV07b]. The authors implemented a method that uses Support Vector Machines combined with NGD, to binary classify search terms. This approach was adjusted to be used in classifying queries according to a corresponding dimension. The main steps are:

- Define the anchor words
- Calculate the NGD value between the training query and the each of the anchor words
- Clean and transform the dataset and remove errors introduced in the retrieval
- Build the n-dimensional training vectors
- Perform parameter tuning using a grid-search and choose the combination with cross-validation
- Divide the dataset into 2/3 for training and 1/3 for test
- Train the model using the trainset
- Evaluate using the testset

The SVM implementation picked is from *e1071* package for the R programming language, based on the *libsvm*. It offers multiple types of classification, kernels, and cross-validation. It was configured with C-classification as the class separation algorithm with a Radial basis function (RBF) kernel. The C and Gamma parameters were tuned using a grid search and combined using cross-validation. The values set for C are {0.1,1,10,100} and Gamma are {0.5,1,2}. We picked C-classification with RBF due to its "good general performance and few number of parameters" [DD01]. It is also important to state that all the classifications were made using Portuguese and

English datasets apart from the semantic types classification due to the nonexistence of labeled queries. In some classifiers, there was the need of performing class balancing. For that, we used the downSample function to randomly balance the frequency of all classes to match the frequency of the minority class.

Chapter 4

Health-related Classifier

This chapter presents two health-related classifiers: NGD Direct Comparison that compares multiple thresholds with the NGD value and an SVM classifier that takes advantage of the NGD.

4.1 NGD Direct Comparison

This first method has the objective to identify a query as health-related or non-health related using the Normalized Google Distance. The method is based on the idea proposed by Eysenbach and Kohler [EK03] that health terms in web queries should co-occur more frequently with the word "health" that non-health terms but instead of determining the co-occurrence rate, we calculate the NGD as shown in 4.1. Apart from the terms "health" we also tested we the terms "medicine" and "health + medicine" and "saude", "medicina" and "saude + medicina" for the Portuguese language. To determine the NGD value we also need to define the M parameter. This was estimated by the most popular preposition in both languages, "the" for English and "a" or "o" for Portuguese. The respective value are 350,000,000 and 256,000,000. After determining the NGD value, we compared it with multiple defined thresholds (0.7, 0.6, 0.5, 0.4, 0.3). If the calculated value is lower than the threshold, then the query is classified as health-related.

$$NGD(terms, health) = \frac{\max\{\log f(terms), \log f(health)\} - \log f(terms, health)}{\log M - \min\{\log f(terms), \log f(health)\}} \tag{4.1}$$

4.1.1 Results

This method was evaluated against the three datasets previously described in section 2.3.3. The NGD values for the KDD dataset do not match the expected range presented by its authors (NGD values between 0 and 1), as we observed values greater than 1 in 391 queries for the term "health", 548 for "medicine" and 579 for "health+medicine". This pattern also present in the AOl dataset, where we can see NGD values over 1 in 12060, 15995, 16893 queries for the "health", "medicine"

and “medicine+health” terms respectively. The last dataset, SAPO, doesn’t have an category label like the others. On the other hand, it has a sum_health attribute that ranges between 0 and 6 according to how many people classified that query as health-related. Due to this characteristic, we run two experiments: one having the sum_health attribute is greater than 2 and other greater than 4. In this dataset has 857, 1047, 1341 queries with NGD value greater than 1 for “saude”, “medicina”, “saude+medicina”.

As seen in the Tables 4.1, 4.2, 4.3, 4.4 precision and recall values are really low when comparing with previous work. In the English datasets, AOL and KDD, both precison and recall metrics are poor, being the “health” term with the lowest performance. On the other hand, the Portuguese dataset has reasonably high precision but low recall.

Table 4.1: Precision (P), Recall (R), F1-score (F1) and Cohen’s Kappa (K) for Kdd dataset

Threshold	**Health**				**Medicine**				**Health+Medicine**			
	P	**R**	**F1**	**K**	**P**	**R**	**F1**	**K**	**P**	**R**	**F1**	**K**
0.7	0.10	0.37	0.15	0.04	0.16	0.33	0.22	0.10	0.17	0.3	0.22	0.10
0.6	0.12	0.35	0.18	0.08	0.17	0.25	0.21	0.09	0.17	0.22	0.19	0.07
0.5	0.13	0.23	0.17	0.07	0.17	0.18	0.18	0.07	0.20	0.18	0.19	0.08
0.4	0.15	0.20	0.17	0.08	0.22	0.16	0.19	0.09	0.30	0.13	0.22	0.11
0.3	0.15	0.20	0.18	0.09	0.30	0.13	0.19	0.10	0.32	0.13	0.19	0.09

Table 4.2: Precision (P), Recall (R), F1-score (F1) and Cohen’s Kappa (K) for AOL dataset

Threshold	**Health**				**Medicine**				**Health+Medicine**			
	P	**R**	**F1**	**K**	**P**	**R**	**F1**	**K**	**P**	**R**	**F1**	**K**
0.7	0.05	0.29	0.09	0.01	0.14	0.44	0.21	0.06	0.17	0.46	0.25	0.09
0.6	0.05	0.29	0.08	0	0.17	0.40	0.24	0.08	0.20	0.42	0.28	0.11
0.5	0.05	0.16	0.08	0	0.20	0.36	0.26	0.10	0.25	0.39	0.31	0.13
0.4	0.06	0.13	0.07	0	0.25	0.34	0.29	0.11	0.32	0.36	0.34	0.15
0.3	0.05	0.11	0.07	0.01	0.31	0.31	0.32	0.13	0.40	0.33	0.36	0.16

Table 4.3: Precision (P), Recall (R), F1-score (F1) and Cohen’s Kappa (K) for SAPO dataset considering sum_health > 2

Threshold	**Saude**				**Medicina**				**Saude+Medicina**			
	P	**R**	**F1**	**K**	**P**	**R**	**F1**	**K**	**P**	**R**	**F1**	**K**
0.7	0.64	0.1	0.20	0	0.65	0.1	0.17	0	0.95	0.05	0.1	0.02
0.6	0.61	0.06	0.12	0	0.67	0.07	0.12	0	0.96	0.05	0.10	0.02
0.5	0.67	0.05	0.11	0	0.73	0.061	0.11	0	0.96	0.05	0.09	0.02
0.4	0.75	0.05	0.10	0	0.85	0.05	0.10	0.01	0.96	0.05	0.09	0.02
0.3	0.85	0.05	0.10	0.01	0.9	0.054	0.10	0.01	0.98	0.05	0.9	0.02

Table 4.4: Precision (P), Recall (R), F1-score (F1) and Cohen's Kappa (K) for SAPO dataset considering sum_health > 4

Threshold	Saude				Medicina				Saude+Medicina			
	P	R	F1	K	P	R	F1	K	P	R	F1	K
0.7	0.44	0.17	0.25	0.03	0.44	0.14	0.21	0.0	0.91	0.17	0.19	0.07
0.6	0.46	0.1	0.17	0	0.52	0.11	0.19	0.02	0.93	0.1	0.19	0.06
0.5	0.56	0.1	0.17	0.02	0.61	0.1	0.18	0.03	0.93	0.1	0.18	0.06
0.4	0.65	0.1	0.17	0.03	0.73	0.1	0.18	0.05	0.95	0.1	0.18	0.06
0.3	0.76	0.09	0.17	0.04	0.83	0.1	0.18	0.05	0.95	0.1	0.18	0.06

4.1.2 Comparison with previous work

Due to the unexpected results seen in the previous section, we decided to determine the co-occurrence rate using the equation 4.2 for the word "health" and compare its results obtained by Lopes et al. [LR14]. In Table 4.5, the column "Current Values" corresponds to the newly calculated values and "Previous Values" with values determined in that paper. We can observe quite different values in both precision and recall. This suggest that the search engines might have changed their internal result counts and this method is no longer viable nowadays.

$$Q(terms, health) = \frac{f(terms, health)}{f(health)} \tag{4.2}$$

Table 4.5: Comparison between co-occurrence rates

Threshold	Current Values		Previous Values	
	Precision	Recall	Precision	Recall
1	0.06	0.20	0.88	0.12
0.9	0.06	0.22	0.87	0.21
0.8	0.06	0.25	0.85	0.36
0.7	0.06	0.27	0.82	0.51
0.6	0.06	0.30	0.77	0.65
0.5	0.06	0.34	0.69	0.76
0.4	0.05	0.40	0.57	0.84
0.3	0.05	0.47	0.44	0.92
0.2	0.05	0.57	0.29	0.94
0.1	0.05	0.68	0.15	0.98

4.2 SVM with NGD

A key point to apply SVM using the NGD is the choice of the anchor terms. In this case, we defined two positively related words and two negative related words to the health terms. In this case, the anchor words chosen were "health" and "medical" and two other words were picked from the

classification done in Cilibrasi et al. [CV07a] paper that are not health-related, “swimmers” and “crime”. For the Portuguese dataset, the anchors were directly translated from English and resulted in the words “saude”,“medicina”,“nadadores” and “crime”. From this point, the 4-dimensional vectors are built, some parameter tuning is made and the model is trained and evaluated.

The first experiment with the AOL dataset without class balancing showed promising results compared with the Health-related Simple method. With Gamma set at 2 and C at 10, the model has a 0.96 precision, 0.99 recall, 0.97 f1-score, 0.36 Cohen’s kappa and Table 4.6 has the confusion matrix. Since the class is unbalanced and the dataset only contains 5% of health queries these results are not that optimal. Many non-health queries are classified correctly but that mainly due to the high number of non-health queries.

To mitigate this problem, a second experiment balances the classes. This new dataset contains 1906 queries; half are health-related queries, and the other health is not. Setting the Gamma to 0.5 and C to 100, the model predicts with precision 0.86, recall 0.85, f1-value 0.85, Cohen’s kappa 0.71 and Table 4.7 as the confusion matrix.

Table 4.6: Confusion Matrix for AOL dataset - Class Unbalance

		True	
		not health	**health**
Predicted	**not health**	5829	234
	health	48	89

Table 4.7: Confusion Matrix for AOL dataset - Class Balance

		True	
		not health	**health**
Predicted	**not health**	281	43
	health	49	262

The third and four experiments are similar to the above, but instead of using the AOL dataset it uses the SAPO dataset. The unbalanced experiment obtained a precision of 0.66, recall of 0.60, f1-value of 0.63 and Cohen’s kappa of 0.37. Table 4.8 represents the model confusion matrix. The balanced experiment obtains a slightly better precision, recall,f1-value, increasing to 0.70, 0.72,0.71 respectively and a kappa of 0.39.

Table 4.8: Confusion Matrix for SAPO dataset - Class Unbalance

		True	
		not health	**health**
Predicted	**not health**	130	67
	health	84	215

Table 4.9: Confusion Matrix for SAPO dataset - Class Balance

Predicted \ True	not health	health
not health	158	65
health	61	142

4.3 Conclusion

The results from the first classifier weren't the expected. For the KDD dataset, the best precision was obtained when using the combination of "health" and "medicine". The recall remained low in all the threshold, never exceeding 0.37. There is also an increase in accuracy and a decrease in recall as the threshold decreases. In the AOL dataset, this pattern has repeated the precision for the "health" term where it remains almost the same through all the thresholds. For the SAPO dataset, we observed better precision, reaching 0.95 in some cases. In contrast, the recall value is quite low. Comparing the two experiments with this dataset, we observed better value precision values in sum_health > 2 and lower recall and the opposite for the sum_health > 4 experiment. In all these cases, we either have a low recall or precision and which might indicate that merely comparing the NGD value with predefined thresholds there are better methods for classifying queries as health-related. With the SVM method, the results change a lot. We obtained a precision of 0.96 and recall of 0.99, but these results do not represent the efficiency of the classifier. With the use of Cohen's kappa, one can analyze the efficiency of the method more easily. The first experiment with AOL dataset showed a kappa value of 0.36 which can be interpreted as poor or fair. When we balanced the dataset with the same number of health and non-health queries the precision and recall values decrease slightly . In this cases the kappa value increases and reaches the interval considered good. There was not as significant change in metrics by doing class balancing as in previous experience but still showed a slight improvement.

To sum up, the NGD Direct Comparison method, presented in Section 4.1, did not present meaningful results. On the other hand, the outcome of the SVM method is auspicious. It is also noticeable that performing class balancing before training the model improves the classifier.

Unfortunately, results can only be compared with previous work done in field of general query classification and not in field of health query classification due to lack of investigation and published results. Comparing this value with the work done by Shen et al. [SPS+05], the results here presented are similar or slighty lower in the SVM methods for the AOL dataset. There classification obtain an average of 0.94 for the f1-score. Ours obtained 0.97 for the unbalance experiment and 0.85 for the balanced. The same is valid for work subsequent to Shen et al. [SPS+05], which obtained better results.

Chapter 5

Severity Classifier

This chapter presents two severity classifiers: NGD Direct Comparison that compares multiple thresholds with the NGD value and an SVM classifier that takes advantage of the NGD.

5.1 Dataset creation

The concept of severity in medical terms is fuzzy and subjective. There isn't a consensus on the exact definition of severity. However, there are several metrics from which severity can be extrapolated, such as:

- Severity of Illness (SOI) - is a medical classification of the patient mental and physical conditions. It is usually measured into minor, moderate, major and extreme;
- Risk of Mortality - is the probability of a patient dying.

There isn't any public dataset available with this kind of metrics, so we had to find another way to obtain the necessary datasets.

For the Portuguese language we used the SAPO dataset and considered a query as severe based on *sum(severe)* attribute. This attribute represents the sum of human labelers that considered the query had severe medical condition associated. It ranges from 0 to 5. If it is bigger than 0, then the query is considered severe otherwise it is not. This new dataset contains 237 severe labeled queries out of 1551 total queries.

The English dataset is based on the World Health Organization report [Org16] on disease burden and mortality estimates. This report contains data on the global estimated number of deaths by cause divided into the 2015, 2010, 2005 and 2000 years, gender and year. From this data, the new dataset was created with two attributes: query and fatality. The query attribute corresponds to the lowest child in a group causes of death from the WHO report excluding some specific cases where the cause is other(s). For example, in the group of Infectious and parasitic diseases, the first

child is Tuberculosis, so this disease is included in the dataset. The following disease is a group of STDs excluding HIV. This group isn't included but its children are with the exception of Other STDs. The fatality attribute is the number of casualties from both genders, any age and the year of 2015 from that specific disease. In the end, this new dataset contains 124 entries and fatality values ranging from 0 to 8,756,006. Two variants were computed based on this data. One where the query is considered severe when the fatality is bigger than the first quartile, of the fatality sample and a second variant from the median. This last dataset is a deliverable of this work.

5.2 NGD Direct Comparison

This method has the objective of classifying the queries as severe or not severe, taking advantage of the Normalized Google Distance. To do so, we first need to define a set of terms both in English and Portuguese. The terms chosen were the following: "dangerous", "severe", "death" and "fatal" for English and "perigoso", "morte", "grave" and "fatal" for Portuguese. These terms were picked by being heavily related with the concept of death and severe. The next step is to calculate the NGD value for each query and each defined term. The last step consists in comparing the retrieved NGD value with the following thresholds: {0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1}. If the NGD value is below the a certain threshold then the health query is considered to be severe. If it is above then the query is considered not severe.

In Table 5.1 and 5.2 are present the precision and recall for the terms defined previously for the SAPO dataset. The method has high recall values, approximately 100% across all the different terms for high threshold values. The precisions are fairly low in all the terms and thresholds with the exception of the term "morte" with a threshold of 0.1 which has the highest precision of 0.5 but a recall less than 1%.

Table 5.1: Precision (P), Recall (R), F1-score(F1) and Cohen's Kappa (K) for SAPO dataset - "Perigoso" and "Morte" terms

Threshold	**Perigoso**				**Morte**			
	P	**R**	**F1**	**K**	**P**	**R**	**F1**	**K**
0.9	0.16	1	0.28	-0.03	0.16	1	0.28	-0.02
0.8	0.16	1	0.28	-0.02	0.17	1	0.28	-0.02
0.7	0.16	1	0.28	-0.02	0.16	0.94	0.28	-0.03
0.6	0.16	0.956	0.28	-0.02	0.15	0.57	0.23	-0.07
0.5	0.16	0.857	0.28	-0.02	0.13	0.33	0.19	-0.10
0.4	0.16	0.404	0.27	-0.03	0.12	0.18	0.14	-0.14
0.3	0.14	0.46	0.22	-0.08	0.14	0.14	0.14	-0.08
0.2	0.14	0.31	0.19	-0.08	0.20	0.09	0.13	-0.03
0.1	0.12	0.19	0.15	-0.10	0.5	0.08	0.14	0.05

Table 5.2: Precision (P), Recall (R), F1-score(F1) and Cohen's Kappa (K) for SAPO dataset - "Grave" and "Fatal" terms

Threshold	Grave				Fatal			
	P	**R**	**F1**	**K**	**P**	**R**	**F1**	**K**
0.9	0.16	1	0.28	-0.024	0.16	0.10	0.28	-0.03
0.8	0.16	0.988	0.28	-0.03	0.17	0.972	0.28	-0.03
0.7	0.17	0.98	0.29	-0.016	0.19	0.833	0.28	-0.03
0.6	0.17	0.698	0.28	-0.01	0.16	0.361	0.22	-0.05
0.5	0.17	0.437	0.24	-0.02	0.17	0.23	0.20	0
0.4	0.15	0.238	0.19	-0.05	0.21	0.194	0.2	0
0.3	0.15	0.147	0.15	-0.06	0.22	0.147	0.180	0
0.2	0.17	0.091	0.12	-0.04	0.26	0.107	0.15	0
0.1	0.28	0.083	0.13	0.01	0.39	0.095	0.15	0.03

For the English dataset, we noticed some correlation between the NGD values and the fatality without applying the method. Allegedly, lower NGD value means the compared terms are more alike in terms of the word meaning. Since, in this case, we have a large number of discrete values and not labels regarding each query, we could plot the Chart 5.1. It represents the variation of the NGD per term and per fatality. To facilitate the visualization of the pattern it was only plotted in 97% of the dataset, excluding four queries. It uses the geom_smooth function from the ggplot2 library in R and a generalized linear model (glm) as a smoothing method. Each point in the line represents the predicted NGD value for queries with the correspondent number of causalities. The grey area around the line is the confidence interval. There is a tendency for the NGD to decrease as the number of casualties increases. The most significant decrease was for the term "severe" and the lowest for the term "dangerous".

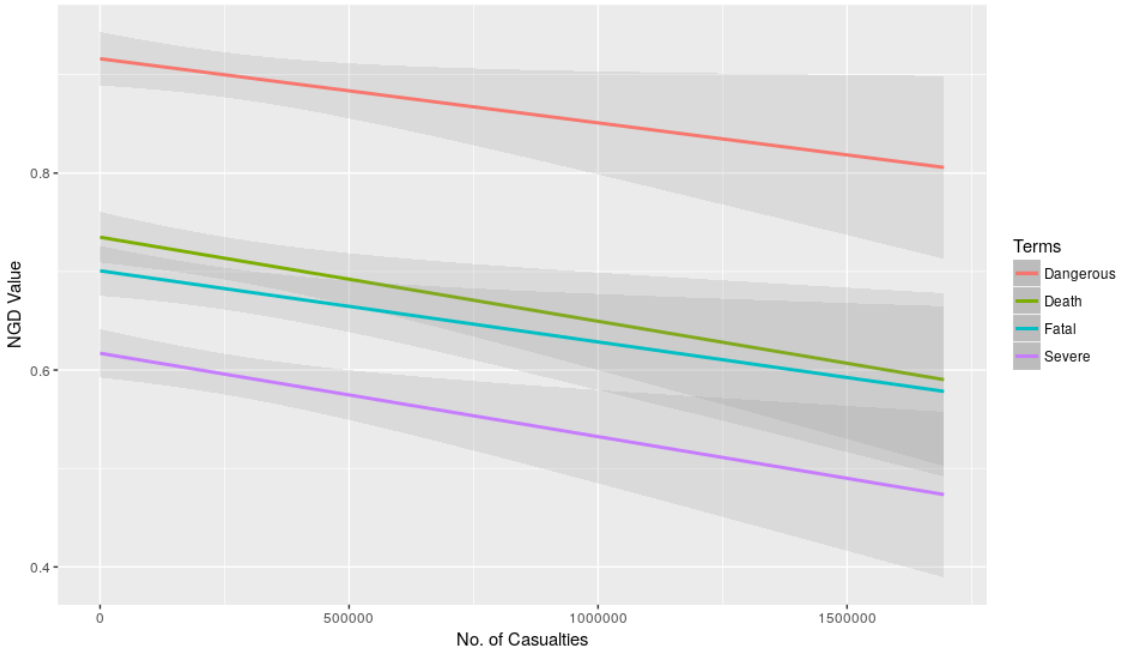

Figure 5.1: NGD values for the terms “dangerous”,“death”,“fatal” and “severe” per no. of Casualties

In Table 5.3, 5.4 and 5.5, 5.6 are the precision and recall values obtain by this method for the EN dataset. The first tables show the results when considering a fatality greater than 6,194, first quartile, as severe queries. In the second table, the value for splitting the dataset between severe and non-severe was the median which has as value 89,877. The cells filled with a dash means the method will not recover anything for that threshold.

The results were the expected. We obtained high recall values in high threshold for both variants except for the word “dangerous” in the median variant. It is also the term with the lowest range of NGD values. Generally the precision values tend to increase sightly in both cases, being the highest in 0.6 and 0.7 threshold for the 1st quartile variant and between 0.7 and 0.4 in the median case. It is also noticed that the world “fatal” achieved 100% precision in both variants. The Cohen’s kappa are fairly low and only reached the value 0.30 three times in the terms “Fatal” and “Death”. Its also noticeable that the interval of NGD values is lower then the SAPO dataset.

Table 5.3: Precision (P), Recall (R), F1-score(F1) and Cohen' Kappa (K) for EN dataset - Median Variant - "Dangerous" and "Death" terms

Threshold	Dangerous				Death			
	P	R	F1	K	P	R	F1	K
0.9	0.60	0.55	0.58	0.21	0.5	1	0.67	0.06
0.8	0.57	0.28	0.37	0.084	0.58	0.85	0.69	0.28
0.7	0.54	0.12	0.20	0.025	0.65	0.667	0.6	0.30
0.6	-	-	-	-	0.5	0.5	0.03	0.07
0.5	-	-	-	-	-	-	-	-
0.4	-	-	-	-	-	-	-	-
0.3	-	-	-	-	-	-	-	-
0.2	-	-	-	-	-	-	-	-
0.1	-	-	-	-	-	-	-	-

Table 5.4: Precision and Recall for EN dataset - Median Variant - "Severe" and "Fatal" terms

Threshold	Severe				Fatal			
	P	R	F1	K	P	R	F1	K
0.9	0.48	1	0.65	0	0.50	0.97	0.66	-0.02
0.8	0.48	0.98	0.65	-0.001	0.52	0.88	0.66	0.08
0.7	0.58	0.93	0.72	0.30	0.64	0.69	0.67	0.31
0.6	0.63	0.59	0.61	0.26	0.76	0.45	0.58	0.32
0.5	0.67	0.28	0.39	0.15	0.78	0.12	0.20	0.08
0.4	0.46	0.09	0.15	-0.011	1	0.03	0.03	0.02
0.3	-	-			-	-		
0.2	-	-			-	-		
0.1	-	-			-	-		

Table 5.5: Precision (P), Recall (R), F1-score(F1) and Cohen' Kappa (K) for EN dataset - 1st Quartile Variant - "Dangerous" and "Death" terms

Threshold	Dangerous				Death			
	P	R	F1	K	P	R	F1	K
0.9	0.80	0.48	0.60	0.09	0.75	0.97	0.85	0
0.8	0.79	0.25	0.38	0.03	0.81	0.76	0.79	0.20
0.7	0.79	0.12	0.21	0.01	0.84	0.51	0.64	0.16
0.6	-	-	-	-	0.77	0.18	0.	0.01
0.5	-	-	-	-	0.5	0.02	0.04	-0.02
0.4	-	-	-	-	-	-	-	-
0.3	-	-	-	-	-	-	-	-
0.2	-	-	-	-	-	-	-	-
0.1	-	-	-	-	-	-	-	-

Table 5.6: Precision (P), Recall (R), F1-score(F1) and Cohen' Kappa (K) for EN dataset - 1st Quartile Variant - "Severe" and "Fatal" terms

Threshold	Severe				Fatal			
	P	R	F1	K	P	R	F1	K
0.9	0.75	1	0.86	0	0.75	0.98	0.85	0.02
0.8	0.75	0.99	0.86	0.03	0.78	0.88	0.83	0.16
0.7	0.79	0.83	0.81	0.19	0.84	0.62	0.72	0.21
0.6	0.82	0.48	0.61	0.115	0.92	0.38	0.53	0.18
0.5	0.8	0.22	0.34	0.03	1	0.10	0.18	0.05
0.4	0.73	0.09	0.15	-0.01	1	0.01	0.02	0.01
0.3	-	-	-	-	-	-	-	-
0.2	-	-	-	-	-	-	-	-
0.1	-	-	-	-	-	-	-	-

5.3 SVM with NGD

Using the Support Vector Machines, the approach was similar to Health-related using SVM method in Section 4.2. Instead of picking 2 random words, the antonyms of the previous terms were used. This way, the model is trained with a 8-dimensional vector, that has as features the NGD values for query compared with the terms: "dangerous", "death", "severe", "fatal", "healthy", "harmless", "mild" and "life" for the English dataset and "perigoso", "morte", "grave", "fatal", "saudavel", "inofensivo", "suave" and "vida". It uses the same ratio for the training data and test data and same grid search with cross-validation for the Gamma and C parameters.

The first experiment resulted in a model with precision of 0.85, recall of 0.99, f1-score of 0.91 and a kappa of -0.015. Given the fact that kappa is approximately 0, the results are considered not adequate and the model isn't classifying correctly the queries. This measure being 0 means the random precision is the same as the observed precision. The Table 5.7 shows the model didn't predict a single query as severe.

In a second attempt, the severe class was balanced. The resultant model was train with less data, the dataset went from 1530 to 464 queries, but showed significantly better results. The precision is now 0.97, recall is 1, f1-score is 0.98, kappa is 0.97 and the confusion matrix corresponds to Table 5.8.

Table 5.7: Confusion Matrix for SAPO dataset - Class Unbalance

		True	
		not severe	**severe**
Predicted	**not severe**	433	73
	severe	3	0

Table 5.8: Confusion Matrix for SAPO dataset - Class Balance

		True	
		not severe	**severe**
Predicted	**not severe**	80	2
	severe	0	72

For the English dataset, we only used the median variant since balancing the classes shows better results and this way we already have 62 severe labeled queries and other 62 non-severe label queries. After training this model with 83 queries and evaluating with the remaining 41 and using as C parameter the value 1 and Gamma the value 0.5, we obtained the confusion matrix of Table 5.9. We noticed a precision of 0.75, recall of 0.78, f1-score of 0.76 and kappa value of 0.53.

Table 5.9: Confusion Matrix for EN dataset

		True	
		not severe	**severe**
Predicted	**not severe**	15	5
	severe	4	17

5.4 Conclusion

In the NGD Direct Comparison method, the dataset in English seen to have better results in both variants than the Portuguese dataset. The "Fatal" term reached 0.32 of Cohen's Kappa, the highest for this method with a precision of 0.76, recall of 0.45 and f1-score of 0.58. This term also had

the highest precision of 100% with a threshold of 0.5 in the median variant and 0.4 in the first quartile variant but low recall. As expected, as the threshold decreases the precision increases and in return, the recall decreases. In the SAPO dataset, the Cohen's kappa is low and reaches negative values in many cases. The increase of precision and decrease is not as accentuated as previously seen but is still present. In general, the method improved a bit when compared to a random classifier for the English dataset set but not for the Portuguese one.

When evaluating the SVM with the Portuguese dataset, we still see slightly negative Cohen's kappa values. Balancing the number of severe and not severe queries resulted in a significant increase in this metric, increasing from -0.015 to 0.97. For the English dataset, there was also an improvement regarding the previous method. The precision remained the same compared to the best result of the previous method, but the recall improved at 0.33 as well as the kappa at 0.21. To conclude, the SVM method obtained better results in both datasets, especially in SAPO and is much more suitable for classifying queries as severe and not severe.

Chapter 6

Semantic Types Classifier

This chapter presents two Semantic Types classifiers: NGD Direct Comparison that compares multiple thresholds with the NGD value in Section 6.3 and an SVM classifier in Section 6.4. The objective of these classifiers is to determine some UMLS Semantic Types present in users' queries. In order to evaluate these methods, we had to construct a dataset, as explained in Section 6.1.

6.1 Dataset Creation

To be able to evaluate these methods, we first need to construct a dataset. The process of creating this dataset was divided into two smaller processes, as shown in Figure 6.1.

The first process consists of obtaining the semantic types associated with each health query in the AOL dataset. It first splits the string into words, then removes the stopwords, next it computes the combinatorics permutations of the query to create a series of subqueries, for example, the query "liver cancer" will produce the following queries: "liver", "cancer", "liver cancer" and "cancer liver". This increased the AOL dataset from 1,198 to 346,578. The next step is querying the UMLS for entities present in each of this newly created queries. This step was very time-consuming. To obtain the entities, it was first necessary to obtain a key named TGT, that is only valid for 8h. Through this key, we obtained the service key that is then used to obtain the ids of the concepts. Finally, a semantic type is extracted by a new request to UMLS. Taking the previous example and assuming that each subquery has five concepts, we have a minimum of 48 requests to the UMLS API. Due to this high number of requests, a limit of 10 concepts per subquery was imposed. The final step is joining the resultant semantic types of the subqueries back to the original one.

The second process has the objective of finding the most frequent terms per semantic types, which will be used later in the NGD calculation. For this, the Consumer Health Vocabulary of 2011 was used. It contains 158,520 terms with a single concept id mapped. Through this id, we

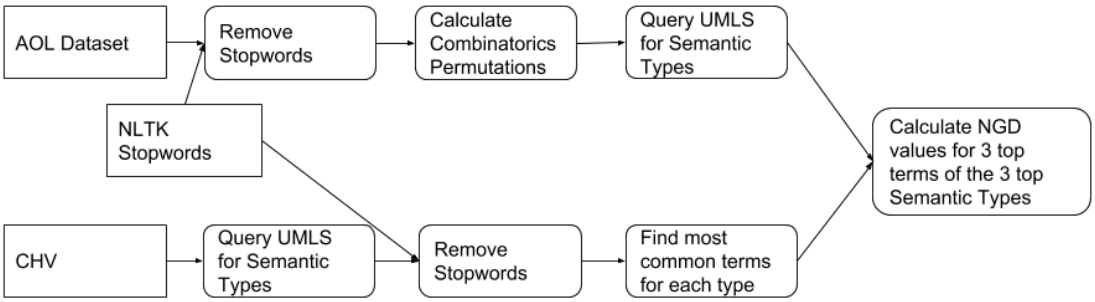

Figure 6.1: Dataset Construction Diagram

get the semantic type, remove all the stopwords present, and estimate the top term for each type. Table 6.1 shows a sample of those results.

Table 6.1: Top 5 terms present in 4 random categories

Finding	Sign or Symptom	Organization	Injury or Poisoning
normal	pain	school	injury
blood	pains	research	fracture
skin	symptoms	center	injuries
abnormal	skin	schools	fractures
urine	chest	care	burn

Finally, the NGD is determined for the top 3 semantic types in the AOL dataset and used as terms of comparison the most frequent ones in the CHV for that specific type. Table 6.2 exhibits the chosen types and terms. This new dataset is a deliverable of this work.

Table 6.2: Chosen Semantic Types and corresponding

Finding	Disease or Syndrome	Therapeutic or Preventive Procedure
normal	syndrome	therapy
blood	disease	surgery
skin	infection	procedure

6.2 Comparing the most searched semantic types with previous results

During the retrieval of semantic types from the generated subqueries, we noticed that the most frequent types present in this queries seem coherent with other studies.

Table 6.3: Top 10 Semantic Types present in subqueries

Semantic Types	Count
Finding	1504
Disease or Syndrome	1087
Therapeutic or Preventive Procedure	959
Intellectual Product	849
Pharmacologic Substance	704
Health Care Activity	672
Organic Chemical	585
Clinical Attribute	445
Amino Acid, Peptide, or Protein	440

To further investigate this suspicion, the top 10 types present in the subqueries, Table 6.3, were compared with results from the "European Citizens' Digital Health Literacy Report" [EUR14], conducted in 2014 and a study conducted in the US in 2011 from the Pew Research Center from Susannah Fox [Fox11].

The first report concluded that 55% of the respondants search for general information on health-related topics or ways to improve their health, 54% on a specific injury, disease, illness or condition, 23% on specific information on medical treatment or procedure, 10% second opinion after a doctor visit and 4% others. Comparing these findings with ours, it is evident that the second and third most searched categories correspond to the ones found in the subqueries. Regarding the most searched category, the UMLS defines the Finding semantic types as "That which is discovered by direct observation or measurement of an organism attribute or condition, including the clinical history of the patient. The history of the presence of a disease is a 'Finding' and is distinguished from the disease itself." and "The result of an examination or inquiry.". This type belongs to a group representing abstract entities which by itself makes Finding an abstract concept. Due to its abstract nature, we cannot directly relate this to a type found in any of these studies, but there might be some correlation with general information on health-related topics present in the E.U. study.

The second report concluded that 66% of internet users search for a specific disease or medical problem, 56% for medical treatment or procedure, 44% on doctors and health professionals, 36% on medical facilities, 33% on health insurance, 29% on food safety and 24% on drug safety. It is again observed the relationship between the first two subjects of the study and the second and third most frequent topic in the subqueries. The type Health Care Activity is also observed in the category of food and drug safety. To sum up, Table 6.4 shows the top 7 types present in our method and the two studies mentioned above. Its evident the relationships between the the types retrieve in our method and the types found in the studies. The most searched types obtained by our automatic method confirm, on a certain level, the discoveries made by these studies.

Table 6.4: Top 7 types of health-related information searched by users

Our Method	**E.U. Report**	**Pew Research**
Finding	General information on health-related topics or ways to improve health	Specific disease or medical problem
Disease or Syndrome	Information on a specific injury, disease, illness or condition	Certain medical treatment or procedure
Therapeutic or Preventive Procedure	Specific information on a medical treatment or procedure	Information on doctors or other health professionals
Intellectual Product	Information to get a second opinion after having visited your doctor	Hospitals or other medical facilities
Pharmacologic Substance	Other	Health Insurance
Health Care Activity	Don't know	Food Safety
Organic Chemical	-	Drug Safety

6.3 NGD Direct Comparison

This method has the objective of determining if the Finding, Disease or Syndrome and Therapeutic or Preventive Procedure semantic types are present in the user queries. These three types were selected due higher presence in queries and relevance to the public and more precisely to health seekers and only as a prove of concept that it is possible to classify an health query into semantic types. It is also important to focus that this is not a multiclass classifier but three binary classifiers, one for each type. Just like the other two threshold classifiers, the NGD is calculated for the three most frequent terms in the CHV for each semantic type as described above. This terms can be found in Table 6.2. To evaluate the quality of the classification, the precision and recall were assessed for each of the following thresholds: {1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1}.

The results for the classification can be found in Tables 6.5, 6.6, 6.7. In a first note, we see very few retrieved values below the 0.3 threshold. There is also a continuous decrease of the recall value as the threshold decreases. The precision values are higher in the middle of the interval of thresholds, mainly between 0.7 and 0.5. The Cohen's Kappa is low, and most of it is classified as poor. There are not many discrepancies in the metrics calculated in the 3 classifiers. The recall values start are approximately at 1, which means that all queries are have an NGD below 1. The accuracy lies between 0.48 and 0.61 and increases by 0.7 and 0.9 for average threshold values.

Table 6.5: Precision (P), Recall (R), F1-score(F1) and Cohen's Kappa (K) for classifying terms of Finding

T	Normal				Blood				Skin			
	P	R	F1	K	P	R	F1	K	P	R	F1	K
1	0.60	0.98	0.74	0.03	0.60	1	0.75	0.03	0.61	0.98	0.75	0.07
0.9	0.64	0.91	0.75	0.15	0.61	0.99	0.76	0.09	0.65	0.86	0.74	0.20
0.8	0.69	0.63	0.66	0.22	0.66	0.93	0.77	0.24	0.71	0.54	0.61	0.20
0.7	0.7	0.33	0.44	0.11	0.69	0.69	0.69	0.24	0.71	0.25	0.37	0.09
0.6	0.65	0.11	0.18	0.02	0.7	0.42	0.52	0.14	0.62	0.08	0.15	0.01
0.5	0.4	0.02	0.04	-0.02	0.72	0.24	0.36	0.09	0.36	0.02	0.04	-0.02
0.4	0.24	0.01	0.02	-0.02	0.67	0.08	0.14	0.02	0.26	0.01	0.01	-0.02
0.3	0.33	0	0.01	-0.01	0.46	0.04	0.04	-0.01	-	-	-	-
0.2	-	-	-	-	-	-	-	-	-	-	-	-
0.1	-	-	-	-	-	-	-	-	-	-	-	-

Table 6.6: Precision (P), Recall (R), F1-score(F1) and Cohen's Kappa (K) for classifying terms of Disease or Syndrome

T	Syndrome				Disease				Infection			
	P	R	F1	K	P	R	F1	K	P	R	F1	K
1	0.54	1	0.70	0.03	0.53	1	0.70	0.03	0.54	0.99	0.70	0.03
0.9	0.54	0.99	0.70	0.06	0.55	0.99	0.70	0.07	0.55	0.97	0.70	0.08
0.8	0.56	0.94	0.70	0.13	0.57	0.92	0.70	0.14	0.58	0.93	0.71	0.18
0.7	0.61	0.83	0.70	0.25	0.60	0.73	0.66	0.19	0.62	0.73	0.67	0.22
0.6	0.64	0.59	0.61	0.21	0.61	0.51	0.56	0.14	0.65	0.51	0.57	0.20
0.5	0.63	0.38	0.48	0.13	0.6	0.29	0.39	0.07	0.65	0.28	0.39	0.11
0.4	0.63	0.21	0.32	0.07	0.52	0.09	0.16	-0.01	0.63	0.12	0.20	0.04
0.3	0.56	0.08	0.14	0.01	0.39	0.03	0.05	-0.02	0.46	0.02	0.04	-0.01
0.2	0.36	0.01	0.03	-0.01	-	-	-	-	0.36	0.01	0.013	-0.01
0.1	0.14	0	0	-0.01	-	-	-	-	-	-	-	-

Table 6.7: Precision (P), Recall (R), F1-score(F1) and Cohen's Kappa (K) for classifying terms of Therapeutic or Preventive Procedure

T	Therapy				Surgery				Procedure			
	P	R	F1	K	P	R	F1	K	P	R	F1	K
1	0.47	0.99	0.64	0.02	0.47	0.99	0.64	0.02	0.48	0.97	0.64	0.02
0.9	0.48	0.97	0.64	0.03	0.48	0.97	0.64	0.05	0.49	0.92	0.64	0.07
0.8	0.51	0.89	0.64	0.11	0.50	0.90	0.64	0.10	0.52	0.78	0.63	0.15
0.7	0.54	0.66	0.59	0.16	0.54	0.71	0.62	0.18	0.56	0.53	0.54	0.16
0.6	0.54	0.40	0.46	0.10	0.56	0.47	0.51	0.14	0.55	0.29	0.38	0.08
0.5	0.58	0.24	0.34	0.09	0.56	0.27	0.37	0.09	0.53	0.09	0.15	0.02
0.4	0.54	0.09	0.15	0.02	0.61	0.13	0.22	0.06	0.35	0.02	0.04	-0.01
0.3	0.42	0.02	0.05	-0.01	0.5	0.03	0.06	0	0.14	0.00	0.01	-0.02
0.2	-	-	-	-	-	-	-	-	0.33	0	0.01	0
0.1	-	-	-	-	-	-	-	-	-	-	-	-

6.4 SVM with NGD

Like the previous method, this consists of 3 distinct binary classifiers, one for each type. The initial idea was to SVM with a multiclass classification variant to determine if the query belongs to Finding, Disease or Syndrome, Therapeutic or Preventive Procedure or to none. This could not be applied because there are many queries that are labeled with more than one of these types. To workaround this issue, we splitted the classifiers into 3 smaller ones, which have the objective of classing the query as a Finding, a Disease or Syndrome or a Therapeutic or Preventive Procedure. It is also noteworthy that a class balancing was performed due to the improvements observed in previous methods.

To determine the negative anchors, the following procedure was adopted:

1. Pick a random semantic type. If this type belongs to the same group as the target type or is a parent or child then another random type is chosen until one meets this condition;
2. The most frequent terms for the chosen type and target type are determined through CHV.
3. These chosen type terms are iterated in decreasing order until it is found one term that does not belong to the semantic type being considered;
4. That term is selected as a negative anchor;
5. Repeat these steps for the remaining anchors.

The Finding classifier was implemented using the positive anchors from Table 6.2 and as negative anchors the terms "malformation", "south" and "bacillus" from the "Congenital Abnormality", "Geographic Area" and "Bacterium". After balancing the classes and removing some

errors, the dataset contained 902 queries. With the cost and gamma parameters set to 1, the classifier obtained a precision of 0.57, recall of 0.59, f1-score of 0.58 and kappa of 0.17. The confusion matrix can be found in Table 6.8.

The Disease or Syndrome classifier also uses the terms in Table 6.2 as positive anchors. For negative anchors, it uses the terms "biopsy", "topical" and "sequences" from the "Diagnostic Procedure", "Biomedical or Dental Material" and "Amino Acid Sequence". The results were slightly better than the Finding classifier. It has a precision of 0.63, recall of 0.64, f1-score of 0.63, kappa of 0.26 and Table 6.6 as confusion matrix. It was tuned with a 0.1 as C and 0.5 as Gamma values.

Lastly, the Therapeutic or Preventive Procedure classifier as the terms "scale", "hybridization" and "archaea" from the "Intellectual Product", "Molecular Biology Research Technique" and "Therapeutic or Preventive Procedure". The resultant model has a precision of 0.59, recall of 0.57, f1-score of 0.58, kappa of 0.19 and as a matrix of confusion the Table 6.10.

Table 6.8: Confusion Matrix for Finding type

		True	
		not finding	**finding**
Predicted	**not finding**	85	64
	finding	59	92

Table 6.9: Confusion Matrix for Disease or Syndrome type

		True	
		not disease	**disease**
Predicted	**not disease**	114	67
	disease	63	110

Table 6.10: Confusion Matrix for Therapeutic or Preventive Procedure type

		True	
		not procedure	**procedure**
Predicted	**not procedure**	104	70
	procedure	78	115

6.5 Conclusion

Before analyzing the results obtained by these two methods, we would like to expose the possibility of errors and misclassifications that may be introduced during the dataset construction phase. For example, there is no guarantee that by splitting the user query into smaller terms and making the combinatorics permutations of them, we are in fact retrieving all the possible UMLS Concept

and consequently the semantic types. Another limitation was only retrieving the first ten concepts per subquery. This restraint had to be imposed by due to API and time limitations. Apart from these problems is very satisfactory to observe that our dataset collect in a automatically way prove the results obtained from the manual inquiries done in Europe and the United States.

Regarding the results from the NGD Direct Comparison method, it is possible to observe, minimal variations in precision and recall between terms from the classification, especially at higher threshold values, which may represent that the terms picked all have similar NGD values for each query. As expected, the recall is almost 100% when the threshold is 1 in all classifications across all terms. Precision is higher between the 0.8 and 0.6 thresholds.

For the SVM method, we noticed a small improvement when comparing to a random classifier. Since the classes are balanced, there is a 50% change for a random classifier to choosing the class the right class for the query. Our classifier improvement of 7% for Finding classification, 13% for Disease or Syndrome and 9% for Therapeutic or Preventive Procedure. These are easily seen with the kappa value. They are between 0.17 and 0.26 which is considered poor by Fleiss et al. [FLP03]. Comparing the results from the two methods for each semantic types, the NGD Direct Comparison method has better results with a threshold of 0.8 for the Finding type with higher precision and recall, kappa and f1-score values but not by much. For Disease or Syndrome, the SVM with NGD classifier obtained similar results when setting the threshold to 0.7 or 0.6. Lastly, the Therapeutic or Preventive Procedure classification using SVM with NGD method obtained slightly higher precision, kappa and f1-core values but lower recall when the threshold for the other method is 0.7. To sum up, the two classifiers had very similar performances, being the SVM with NGD method better for classifying queries in Therapeutic or Preventive Procedure type and NGD Direct Comparison for Finding type.

Chapter 7

Conclusions and Future Work

This chapter presents the goals accomplished in the investigation and a plan for future work on this matter.

7.1 Conclusions

The Web is used actively use by people seeking all kinds of information. There are a number of problems on the way people are searching for this information that makes the retrieval process not as efficient as could be. It can be hard to identify why the user is searching for a determined topic and in what context was it searched. Query Classification has been extensively studied for this purpose.

This dissertation had the objective of studying the feasibility of a semantic similarity metric named Normalized Google Distance in classifying queries. The investigation focused on the health domain and health-related searches. We proposed two types of classifiers, NGD Direct Comparison and SVM with NGD that were applied to determining if a user query was meant to be searching for health information, the severity associated with that health query and in which semantic type was included. The results were satisfactory, and the objectives defined at the beginning of this project were successfully achieved. The SVM with NGD proven to be a better method for performing query classification than the NGD Direct Comparison in the first two dimensions: Health-related and Severity. For the Semantic types both methods achieved similar results. In the end, we conclude that the NGD is a valuable asset in query classification and can be used to improve the context behind a user query.

7.2 Future Work

In the future, and to further investigate the use of this metric in this field, other types of classification could be implemented and improved. Our semantic type classifier could be extended

for classifying more types and tested against a dataset with all the possible types present in the UMLS. A different kind of classification that can be explored is the medical specialty. It was initially proposed but due to lack dataset and ways of constructing way it could not be done.

References

[ABC13] Mohannad ALMasri, Catherine Berrut, and Jean-Pierre Chevallet. Wikipedia-based semantic query enrichment. *Proc. sixth Int. Work. Exploit. Semant. Annot. Inf. Retr. - ESAIR '13*, pages 5–8, 2013.

[AYKZ11] Ritesh Agrawal, Xiaofeng Yu, Irwin King, and Remi Zajac. Enrichment and reductionism: Two approaches for web query classification. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 7064 LNCS(PART 3):148–157, 2011.

[BJF+] S.M. Beitzel, E.C. Jensen, O. Frieder, D.D. Lewis, A. Chowdhury, and A. Kolcz. Improving Automatic Query Classification via Semi-Supervised Learning. In *Fifth IEEE Int. Conf. Data Min.*, pages 42–49. IEEE.

[BJL+07] Steven M. Beitzel, Eric C. Jensen, David D. Lewis, Abdur Chowdhury, and Ophir Frieder. Automatic classification of Web queries using very large unlabeled query logs. *ACM Trans. Inf. Syst.*, 25(2):9–es, 2007.

[BL07] John A. Bullinaria and Joseph P. Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behav. Res. Methods*, 39(3):510–526, aug 2007.

[CHS+] Huanhuan Cao, Derek Hao Hu, Dou Shen, Daxin Jiang, Jian-Tao Sun, Enhong Chen, and Qiang Yang. Context-Aware Query Classification.

[CV07a] Rudi L. Cilibrasi and Paul M B Vitányi. Automatic Meaning Discovery Using Google. *IEEE Trans. Knowl. Data Eng.*, 19(3):370–383, 2007.

[CV07b] Rudi L. Cilibrasi and Paul M.B. Vitányi. The Google similarity distance. *IEEE Trans. Knowl. Data Eng.*, 19(3):370–383, 2007.

[DD01] by FH David Meyer Technikum Wien and Austria DavidMeyer. Support Vector Machines * The Interface to libsvm in package e1071. 13(9), 2001.

[DV09] Eustache Diemert and Gilles Vandelle. Unsupervised query categorization using automatically-built concept graphs. *Proc. 18th Int. Conf. World wide web*, pages 461–470, 2009.

[EK03] G. Eysenbach and Ch. Kohler. What is the prevalence of health-related searches on the World Wide Web? Qualitative and quantitative analysis of search engine queries on the internet. *AMIA Symp. Proc.*, pages 225–9, 2003.

[EKH09] Alberto J Evangelista and Bjørn Kjos-Hanssen. Google Distance Between Words. 2009.

REFERENCES

[EUR14] European citizens' digital health literacy. Technical report, 2014.

[FLP03] Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical Methods for Rates and Proportions*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA, sep 2003.

[Fox11] Susannah Fox. Health Topics. 2011.

[GKG] Knowledge – Inside Search – Google.

[GMM03] R Guna, Rob Mccool, and Eric Miller. Semantic search. *Proc. twelfth Int. Conf. World Wide Web - WWW '03*, pages 700–709, 2003.

[GMX+09] James Geller, C Paul Morrey, Junchuan Xu, Michael Halper, Gai Elhanan, Yehoshua Perl, and George Hripcsak. Comparing inconsistent relationship configurations indicating UMLS errors. *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, 2009:193–7, nov 2009.

[Gri] Breakthrough Analysis: Two + Nine Types of Semantic Search - InformationWeek.

[GtKAvH07] Risto Gligorov, Warner ten Kate, Zharko Aleksovski, and Frank van Harmelen. Using Google distance to weight approximate ontology matches. *Proc. 16th Int. Conf. World Wide Web - WWW '07*, page 767, 2007.

[HSB] Orla Higgins, Jane Sixsmith, and Margaret M Barry. A literature review on health information- seeking behaviour on the web: a health consumer and health professional perspective A literature review on health information-seeking behaviour on the web.

[KSD+08] Alla Keselman, Catherine Arnott Smith, Guy Divita, Hyeoneui Kim, Allen C Browne, Gondy Leroy, and Qing Zeng-Treitler. Consumer health concepts that do not map to the UMLS: where do they fit? *J. Am. Med. Inform. Assoc.*, 15(4):496–505, 2008.

[LB12] Dieu-Thu Le and Raffaella Bernardi. Query classification using topic models and support vector machine. *Proc. ACL 2012 Student Res. Work.*, (July):19–24, 2012.

[LK77] J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159, mar 1977.

[LR14] Carla Teixeira Lopes and Cristina Ribeiro. Identification and Classification of Health Queries. *Int. J. Healthc. Inf. Syst. Informatics*, 9(3):55–71, 2014.

[MRS09] Christopher D. Manning, Prabhakar Ragahvan, and Hinrich Schutze. An Introduction to Information Retrieval. *Inf. Retr. Boston.*, (c):1–18, 2009.

[NM04] Fabiane Bizinella Nardon and Lincoln A Moura. Knowledge sharing and information integration in healthcare using ontologies and deductive databases. *Stud. Health Technol. Inform.*, 107(Pt 1):62–6, 2004.

[Org16] World Health Organization. Global Health Estimates 2015: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2015. Technical report, Geneva, 2016.

REFERENCES

[PDC+] Bernardo Pereira Nunes, Stefan Dietze, Marco Antonio Casanova, Ricardo Kawase, Besnik Fetahu, and Wolfgang Nejdl. Combining a co-occurrence-based and a semantic measure for entity linking.

[PMS+01] T B Patrick, H K Monga, M E Sievert, J Houston Hall, and D R Longo. Evaluation of controlled vocabulary resources for development of a consumer entry vocabulary for diabetes. *J. Med. Internet Res.*, 3(3):E24, 2001.

[PZ06] Jan Poland and Thomas Zeugmann. Clustering the Google Distance with Eigenvectors and Semidefinite Programming. *Knowl. Media Technol. First Int. Core-to-Core Work.*, 21:61–69, 2006.

[sep] What are the top 10 most popular search engines? | Search Engine Watch.

[SPS+05] Dou Shen, Rong Pan, Jian-Tao Sun, Jeffrey Junfeng Pan, Kangheng Wu, Jie Yin, and Qiang Yang. Q 2 C@UST: Our Winning Solution to Query Classification in KDDCUP 2005. *ACM SIGKDD Explor. Newsletter,*, 7(2):100–110., 2005.

[SPS+06] Dou Shen, Rong Pan, Jian-Tao Sun, Jeffrey Junfeng Pan, Kangheng Wu, Jie Yin, and Qiang Yang. Query enrichment for web-query classification. *ACM Trans. Inf. Syst.*, 24(3):320–352, 2006.

[SSYC06] Dou Shen, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Building bridges for web query classification. *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pages 131–138, 2006.

[SWJS] Amanda Spink, Dietmar Wolfram, Major B J Jansen, and Tefko Saracevic. Searching the Web: The Public and Their Queries.

[SYJ+04] Amanda Spink, Yin Yang, Jim Jansen, Pirrko Nykanen, Daniel P. Lorence, Seda Ozmutlu, and H. Cenk Ozmutlu. A study of medical and health queries to web search engines. *Heal. Inf. Libr. J.*, 21(1):44–51, mar 2004.

[TL07] Elaine G. Toms and Celeste Latter. How consumers search for health information. *Health Informatics J.*, 13(3):223–235, 2007.

[uml09] *UMLS® Reference Manual*. National Library of Medicine (US), 2009.

[VMHZ14] V G Vinod Vydiswaran, Qiaozhu Mei, David A Hanauer, and Kai Zheng. Mining consumer health vocabulary from community-generated text. *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, 2014:1150–9, 2014.

[WH09] Ryen W White and Eric Horvitz. Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search. *ACM Trans. Inf. Syst*, 27(23), 2009.

[YHH15] Hebin Yang, Qinmin Hu, and Liang He. Learning Topic-Oriented Word Embedding for Query Classification. In Tru Cao, Ee-Peng Lim, Zhi-Hua Zhou, Tu-Bao Ho, David Cheung, and Hiroshi Motoda, editors, *Adv. Knowl. Discov. Data Min.*, pages 188–198, Cham, 2015. Springer International Publishing.

[ZT06] Q Zeng and T Tse. Exploring and developing consuming health vocabulary. *J Am Med Inf. Assoc*, 13(1):24–29, 2006.

REFERENCES

[ZTC+05] Qing T Zeng, Tony Tse, Jon Crowell, Guy Divita, Laura Roth, and Allen C Browne. Identifying consumer-friendly display (CFD) names for health concepts. *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, 2005:859–63, 2005.